

# The SWISS-MODEL Repository and associated resources

Florian Kiefer<sup>1,2</sup>, Konstantin Arnold<sup>1,2</sup>, Michael Künzli<sup>1,2</sup>, Lorenza Bordoli<sup>1,2</sup>  
and Torsten Schwede<sup>1,2,\*</sup>

<sup>1</sup>Biozentrum, University of Basel and <sup>2</sup>SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Received September 15, 2008; Accepted October 5, 2008

## ABSTRACT

**SWISS-MODEL Repository** (<http://swissmodel.expasy.org/repository/>) is a database of 3D protein structure models generated by the SWISS-MODEL homology-modelling pipeline. The aim of the SWISS-MODEL Repository is to provide access to an up-to-date collection of annotated 3D protein models generated by automated homology modelling for all sequences in Swiss-Prot and for relevant models organisms. Regular updates ensure that target coverage is complete, that models are built using the most recent sequence and template structure databases, and that improvements in the underlying modelling pipeline are fully utilised. As of September 2008, the database contains 3.4 million entries for 2.7 million different protein sequences from the UniProt database. SWISS-MODEL Repository allows the users to assess the quality of the models in the database, search for alternative template structures, and to build models interactively via SWISS-MODEL Workspace (<http://swissmodel.expasy.org/workspace/>). Annotation of models with functional information and cross-linking with other databases such as the Protein Model Portal (<http://www.proteinmodelportal.org>) of the PSI Structural Genomics Knowledge Base facilitates the navigation between protein sequence and structure resources.

## INTRODUCTION

Three-dimensional protein structures are crucial for understanding protein function at a molecular level. In recent years, tremendous progress in experimental techniques for large-scale protein structure determination by X-ray crystallography and NMR has been achieved. Structural genomic efforts have contributed significantly

to the elucidation of novel protein structures (1), and to the development of technologies, which have increased the speed and success rate at which structures can be determined and lowered the cost of the experiments (2,3). However, the number of known protein sequences grows at an ever higher rate as large-scale sequencing projects, such as the Global Ocean Sampling expedition, are producing sequence data at an unprecedented rate (4). Consequently, the last release of the UniProt (5) protein knowledgebase (version 14.0) contained more than 6.5 million sequences, which is about 100 times the number protein structures currently deposited in the Protein Data Bank (6) (~53 000, September 2008). For the foreseeable future, stable and reliable computational approaches for protein structure modelling will therefore be required to derive structural information for the majority of proteins, and a broad variety of *in silico* methods for protein structure prediction has been developed in recent years.

Homology (or comparative) modelling techniques have been shown to provide the most accurate models in cases, where experimental structures related to the protein of interest were available. Although the number of protein sequence families increases at a rate that is linear or almost linear with the addition of new sequences (4), the number of distinct protein folds in nature is limited (1,7) and the growth in the complexity of protein families appears as a result of the combination of domains (M. Levitt, manuscript in preparation). Achieving complete structural coverage of whole proteomes (on the level of individual soluble domain structures) by combining experimental and comparative modelling techniques therefore appears to be a realistic goal, and is already been pursued, e.g. by the Joint Center for Structural Genomics for the small model organism *Thermotoga maritima* (JCSG) (8,9). Assessment of the accuracy of methods for protein structure prediction, e.g. during the bi-annual CASP (Critical Assessment of Techniques for Protein Structure Prediction) experiments (10,11) or the automated EVA project (12), has demonstrated that comparative protein structure modelling is currently the most accurate

\*To whom correspondence should be addressed. Tel: +41 61 267 15 81; Fax: +41 61 267 15 84; Email: [torsten.schwede@unibas.ch](mailto:torsten.schwede@unibas.ch)

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© 2008 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

technique for prediction of the 3D structure of proteins. During the CASP7 experiment, it became apparent that the best fully automated modelling methods have improved to a level where they challenge most human predictors in producing the most accurate models (13–15). Nowadays, comparative protein structure models are often sufficiently accurate to be employed for a wide spectrum of biomedical applications, such as structure based drug design (16–20), functional characterization of diverse members of a protein family (21), or rational protein engineering, e.g. the humanization of therapeutic antibodies, or to study functional properties of proteins (22–26).

Here, we describe the SWISS-MODEL Repository, a database of annotated protein structure models generated by the SWISS-MODEL Pipeline, and a set of associated web-based services that facilitate protein structure modelling and assessment. We emphasize the improvements of the SWISS-MODEL Repository which have been implemented since our last report (27). These include a new pipeline for template selection, the integration with interactive tools in the SWISS-MODEL Workspace, the programmatic access via DAS (distributed annotation system) (28), the implementation of a reference frame for protein sequences based on md5 cryptographic hashes, and the integration with the Protein Model Portal (<http://www.proteinmodelportal.org>) of the PSI Structural Genomics Knowledge Base (29,30).

## REPOSITORY CONTENTS, ACCESS AND INTERFACE

### Homology modelling

The SWISS-MODEL Repository contains models that are calculated using a fully automated homology modelling pipeline. Homology modelling typically consists of the following steps: selection of a suitable template, alignment of target sequence and template structure, model building, energy minimization and/or refinement and model quality assessment. This requires a set of specialized software tools as well as up-to-date sequence and structure databases. The SWISS-MODEL pipeline (version 8.9) integrates these steps into a fully automated workflow by combining the required programs in a PERL based framework.

Since template search and selection is a crucial step for successful model building, we have implemented a hierarchical template search and selection protocol, which is sufficiently fast to be used for automated large-scale modelling, sensitive in detecting low homology targets, and accurate in correctly identifying close target structures. In the first step, segments of the target sequence sharing close similarity to known protein structures are identified using a conservative BLAST (31) search with restrictive parameters [*E*-value cut-off:  $10^{-5}$ , 60% minimum sequence identity to sequences of the SWISS-MODEL Template Library SMTL (32)]. This ensures that information about close sequence relationships is not dispersed by the subsequent profile-based search strategies (33). If regions of the target sequence remain

uncovered, in the second step a search for suitable templates is performed against a library of Hidden Markov Models for SMTL using HHSearch (14). Templates resulting from both steps are ranked according to their *E*-value, sequence identity to the target, resolution and structure quality. From this ranked list, the best templates are progressively selected to maximize the length of the modelled region of the protein. New templates are added if they significantly increase the coverage of the target sequence (spanning at least 25 consecutive residues), or new information is gained (e.g. templates spanning several domains help to infer relative domain orientation). For each selected target–template alignment, 3D models are calculated using ProModII (34) and energy minimized using the Gromos force field (35). The quality of the resulting model is assessed using the ANOLEA mean force potential (36).

Depending on the size of the protein and the evolutionary distance to the template, model building can be relatively time-consuming. Therefore, comprehensive databases of pre-computed models (27,37,38) have been developed in order to be able to cross-link real-time model information with other biological data resources, such as sequence databases or genome browsers.

### Model database

The SWISS-MODEL Repository is a relational database of models generated by the automated SWISS-MODEL pipeline based on protein sequences from the UniProt database (5). Within the database, model target sequences are uniquely identified by their md5 cryptographic hash of the full length raw amino acid sequence. This mechanism allows the redundancy in protein sequence databases entries to be reduced, and facilitates cross-referencing with databases using different accession code systems. Mapping between UniProt and various database accession code systems to our md5 based reference system is derived from the iProClass database (39). Regular updates are performed for all protein sequences in the SwissProt database (40), as well as complete proteomes of several model organisms (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Hepacivirus*). Incremental updates are performed on a regular basis in order to both include new target sequences from the UniProt database and to take advantage of newly available template structures, whereas full updates are required when major improvements to the underlying modelling algorithms have been made. The current SWISS-MODEL Repository release contains 3.45 million models for 2.72 million unique sequences, built on 26 185 different template structures (34 540 chains), covering 48.8% of the entries from UniProt (14.0), and more specifically 65.4% of the unique sequences of Swiss-Prot (56.0), the manually annotated section of the UniProt knowledgebase. The size of the models ranges from 25 up to 2059 residues (e.g. fatty acid synthase  $\beta$ -subunit from *Thermomyces lanuginosus*) with an average model length of 221 residues.

### Graphical user web interface

The web interface at <http://swissmodel.expasy.org/repository/> provides the main entry point to the SWISS-MODEL Repository. Models for specific proteins can be queried using different database accession codes (e.g. UniProt AC and ID, GenBank, IPI, Refseq) or directly with the protein amino acid sequence (or fragments thereof, e.g. for a specific domain). For a given target protein, a graphical overview illustrating the segments for which models (or experimental structures) are available is shown (Figure 1). Functional and domain annotation for the target protein is retrieved dynamically in real time using web service protocols to ensure that the annotation information is up-to-date. UniProt annotation of the target protein is retrieved via REST queries (<http://www.uniprot.org>). Structural domains in the target protein are annotated by PFAM domain assignment (41), which is retrieved dynamically by querying the InterPro (42) database using the DAS protocol (28). The md5-based reference frame for target proteins allows to update the database accession mappings in between modelling release cycles. This ensures that cross references with functional annotation resources such as InterPro correspond to proteins of identical primary sequence, thereby avoiding commonly observed problems with incorrect cross-references as a result of instable accession codes or asynchronous updates of different data resources. Finally, for each model, a summary page provides information on the modelling process (template selection and alignment), model quality assessment by ANOLEA (36) and Gromos (35), and in page visualization of the structure using the Astex Viewer (43) plugin.

### Integration with SWISS-MODEL Workspace

The SWISS-MODEL Repository is a large-scale database of pre-computed 3D models. Often however, one may be interested in performing additional analyses either on the models themselves, or on the underlying protein target sequence. We have therefore implemented a tight link between the entries of the SWISS-MODEL Repository and the corresponding modules in the SWISS-MODEL Workspace, which provides an interactive web-based, personalized working environment (32,34,44). Besides the functionality for building protein models it provides various modules to assess protein structures and models. The estimation of the quality of a protein model is an important step to assess its usefulness for specific applications. In particular, models based on template structures sharing low sequence identity require careful evaluation. Therefore, entries from the Repository can be directly submitted to the Workspace for quality assessment using different global and local quality scores such as DFire (45), ProQRes (46) or QMEAN (47).

The default output format for models in the Repository is the project file for the program DeepView (34); this program allows the underlying alignments to be adjusted manually and for the request to be resubmitted to Workspace for modelling. While new protein structures are deposited in the PDB on a daily basis, the respective modelling update cycles are more infrequent, resulting in

a delay in the incorporation of new templates. The Repository therefore links directly to the corresponding template search module in Workspace, which allows searches for newly released templates to be performed. The direct cross-linking between Repository and Workspace allows combining the advantages of the database of pre-computed models with the flexibility of an interactive modelling system.

## INTEROPERABILITY

### Programmatic access

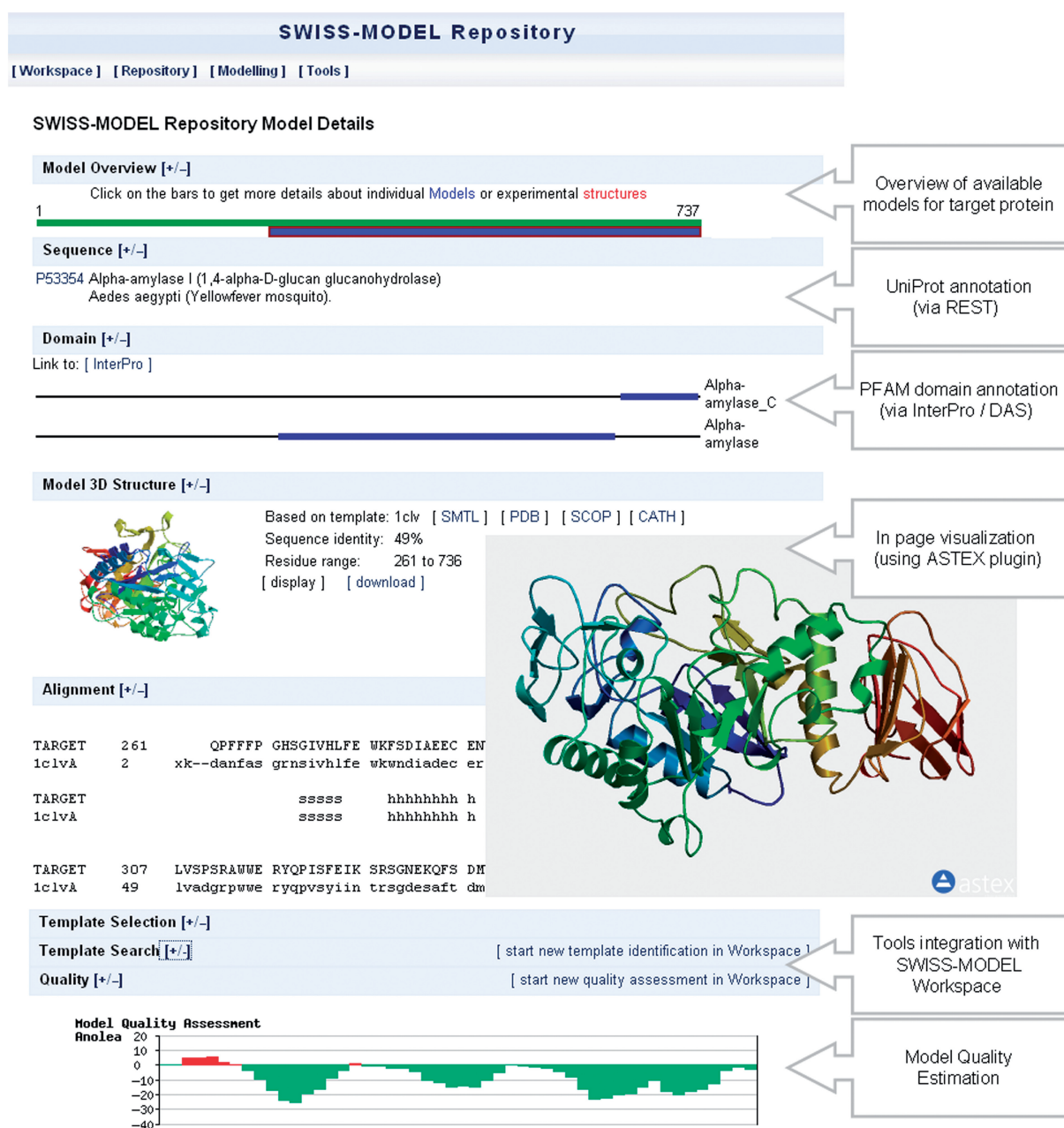
One of the major challenges of computational biology today is the integration of large amounts of diverse data in heterogeneous formats. Very often, data exchange within one domain, e.g. sequence-based data resources, is relatively straightforward, but seamless exchange between resources serving different data types, such as genome browsers and protein structure databases, is more difficult due to the lack of common and accepted standards. DAS (28) is a light-weight mechanism for web service-based annotation exchange. The DAS concept relies on a XML specification which defines the communication between server and client. Queries can be executed by sending a specific http-request to the DAS server. The result of the DAS-Server request is a human readable and easy-to-parse XML-document following the Biodas specifications (<http://www.biodas.org>).

The DAS-Server of the SWISS-MODEL Repository is based on the DAS/1 standard and can be queried by primary UniProt accession codes or md5-hashes of the corresponding sequences. Individual models for a query sequence ('SEGMENT') are annotated as 'FEATURE', with information about the start and stop position in the target sequence, template-sequence identity and the URL to the corresponding SWISS-MODEL Repository entry. The DAS service allows the SWISS-MODEL Repository to be cross-linked with other resources using the same standards, e.g. genome browsers. The SWISS-MODEL Repository DAS service is accessible at <http://swissmodel.expasy.org/service/das/swissmodel/>.

### The protein model portal

One of the major bottlenecks in the use of protein models is that, unlike for experimental structures, modelling resources are heterogeneous and distributed over numerous servers. However, it is often beneficial for the user to directly compare the results of different modelling methods for the same protein. We have therefore developed the protein model portal (PMP) as a component of the PSI structural genomics knowledge base (29,30). This resource provides access to all structures in the PDB, functional annotations, homology models, structural genomics protein target tracking information, available protocols and the potential to obtain DNA materials for many of the targets. The PMP currently provides access to several million pre-built models from four PSI centers, ModBase (38) and SWISS-MODEL Repository (27,37).





**Figure 1.** Typical view of a SWISS-MODEL Repository entry. For the UniProt entry P53354, the  $\alpha$ -amylase I (EC 3.2.1.1; 1,4- $\alpha$ -D-glucan glucanohydrolase) from *Aedes aegypti* (Yellowfever mosquito), a model covering the active amylase domain is shown, including information on the template structure used for model building, the target–template sequence alignment, and quality assessment of the model. Functional annotation such as PFAM domain structure and UniProt annotation of the protein sequence is retrieved dynamically. Links to SWISS-MODEL Workspace enable the user to run additional model quality assessment tools on the model, or search the template library for alternative template structures.

## FUTURE DIRECTIONS

SWISS-MODEL Repository will be updated regularly to reflect the growth of the sequence and structure databases. Future releases of SWISS-MODEL Repository will include models of oligomeric assemblies, as well as models including essential co-factors, metal ions and structural ligands. Structural clustering of the Swiss Model Template Library will also allow us to routinely include ensembles of models for such proteins, which undergo extensive domain movements.

## CITATION

Users of SWISS-MODEL Repository are requested to cite this article in their publications.

## ACKNOWLEDGEMENTS

We are grateful to Rainer Pöhlmann, [BC]<sup>2</sup> & Biozentrum University of Basel, for professional systems support, Pascal Benkert for fruitful discussions on model quality assessment and Jürgen Kopp for pioneering work on

earlier versions of SWISS-MODEL Repository. We thank James Battey for critically reading the article. We are indebted to Dr Michael Podvynec for his enthusiastic support and excellent coordination of the Scrum process for the SWISS-MODEL team. We are grateful to Eric Jain for the swift implementation of md5 based REST queries on the UniProt server, and Wendy Tao, John Westbrook and Helen Berman (RCSB) for the great collaboration on the PSI SGKB Protein Model Portal. Computational resources for SWISS-MODEL Repository are provided by [BC]<sup>2</sup> Basel Computational Biology Center (<http://www.bc2.ch>) and Vital-IT (<http://www.vital-it.ch>).

## FUNDING

The PSI SGKB Protein Model Portal was supported by the National Institutes of Health as a sub-grant with Fox Chase Cancer Center (3 P20 GM076222-02S1); as a sub-grant with Rutgers University, under Prime Agreement Award Number (3U54GM074958-04S2). SWISS-MODEL Workspace and Repository have been supported by the Swiss Institute of Bioinformatics (SIB). Funding for open access charges: Swiss Institute of Bioinformatics.

*Conflict of interest statement.* None declared.

## REFERENCES

- Levitt, M. (2007) Growth of novel protein structural data. *Proc. Natl Acad. Sci. USA*, **104**, 3183–3188.
- Slabinski, L., Jaroszewski, L., Rodrigues, A.P., Rychlewski, L., Wilson, I.A., Lesley, S.A. and Godzik, A. (2007) The challenge of protein structure determination—lessons from structural genomics. *Protein Sci.*, **16**, 2472–2482.
- Manjasetty, B.A., Turnbull, A.P., Panjekar, S., Bussow, K. and Chance, M.R. (2008) Automated technologies and novel techniques to accelerate protein crystallography for structural genomics. *Proteomics*, **8**, 612–625.
- Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W. et al. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.*, **5**, e16.
- Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. et al. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
- Berman, H., Henrick, K., Nakamura, H. and Markley, J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
- Chothia, C. (1992) Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543–544.
- McCleverty, C.J., Columbus, L., Kreusch, A. and Lesley, S.A. (2008) Structure and ligand binding of the soluble domain of a *Thermotoga maritima* membrane protein of unknown function TM1634. *Protein Sci.*, **17**, 869–877.
- Xu, Q., Kozbial, P., McMullan, D., Krishna, S.S., Brittain, S.M., Ficarro, S.B., DiDonato, M., Miller, M.D., Abdubek, P., Axelrod, H.L. et al. (2008) Crystal structure of an ADP-ribosylated protein with a cytidine deaminase-like fold, but unknown function (TM1506), from *Thermotoga maritima* at 2.70 Å resolution. *Proteins*, **71**, 1546–1552.
- Kopp, J., Bordoli, L., Battey, J.N., Kiefer, F. and Schwede, T. (2007) Assessment of CASP7 predictions for template-based modeling targets. *Proteins*, **69**(Suppl 8), 38–56.
- Kryshtafovych, A., Fidelis, K. and Moulton, J. (2007) Progress from CASP6 to CASP7. *Proteins*, **69**(Suppl 8), 194–207.
- Koh, I.Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A. et al. (2003) EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res.*, **31**, 3311–3315.
- Battey, J.N., Kopp, J., Bordoli, L., Read, R.J., Clarke, N.D. and Schwede, T. (2007) Automated server predictions in CASP7. *Proteins*, **69**(Suppl 8), 68–82.
- Soding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
- Zhang, Y. (2007) Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins*, **69**(Suppl 8), 108–117.
- Hillisch, A., Pineda, L.F. and Hilgenfeld, R. (2004) Utility of homology models in the drug discovery process. *Drug Discov. Today*, **9**, 659–669.
- Tan, E.S., Groban, E.S., Jacobson, M.P. and Scanlan, T.S. (2008) Toward deciphering the code to aminergic G protein-coupled receptor drug design. *Chem. Biol.*, **15**, 343–353.
- Thorsteinsdottir, H.B., Schwede, T., Zoete, V. and Meuwly, M. (2006) How inaccuracies in protein structure models affect estimates of protein-ligand interactions: computational analysis of HIV-1 protease inhibitor binding. *Proteins*, **65**, 407–423.
- Vangrevelinghe, E., Zimmermann, K., Schoepfer, J., Portmann, R., Fabbro, D. and Furet, P. (2003) Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J. Med. Chem.*, **46**, 2656–2662.
- Oshiro, C., Bradley, E.K., Eksterowicz, J., Evensen, E., Lamb, M.L., Lancot, J.K., Putta, S., Stanton, R. and Grootenhuys, P.D. (2004) Performance of 3D-database molecular docking studies into homology models. *J. Med. Chem.*, **47**, 764–767.
- Murray, P.S., Li, Z., Wang, J., Tang, C.L., Honig, B. and Murray, D. (2005) Retroviral matrix domains share electrostatic homology: models for membrane binding function throughout the viral life cycle. *Structure*, **13**, 1521–1531.
- Lippow, S.M., Wittrup, K.D. and Tidor, B. (2007) Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.*, **25**, 1171–1176.
- Junne, T., Schwede, T., Goder, V. and Spiess, M. (2006) The plug domain of yeast Sec61p is important for efficient protein translocation, but is not essential for cell viability. *Mol. Biol. Cell*, **17**, 4063–4068.
- Peitsch, M.C. (2002) About the use of protein models. *Bioinformatics*, **18**, 934–938.
- Tramontano, A. (2008) The biological applications of protein models. In Schwede, T. and Peitsch, M.C. (eds), *Computational Structural Biology*. World Scientific Publishing, Singapore.
- Li, Y., Drummond, D.A., Sawayama, A.M., Snow, C.D., Bloom, J.D. and Arnold, F.H. (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.*, **25**, 1051–1056.
- Kopp, J. and Schwede, T. (2006) The SWISS-MODEL Repository: new features and functionalities. *Nucleic Acids Res.*, **34**, D315–D318.
- Jenkinson, A.M., Albrecht, M., Birney, E., Blankenburg, H., Down, T., Finn, R.D., Hermjakob, H., Hubbard, T.J., Jimenez, R.C., Jones, P. et al. (2008) Integrating biological data – the distributed annotation system. *BMC Bioinformatics*, **9**(Suppl 8), S3.
- Berman, H.M., Westbrook, J.D., Gabanyi, M.J., Tao, Y., Shah, R., Kouranov, A., Schwede, T., Arnold, K., Kiefer, F., Bordoli, L. et al. (2008) PSI structural genomics knowledge base. *Nucleic Acids Res.*, in press.
- Berman, H.M. (2008) Harnessing knowledge from structural genomics. *Structure*, **16**, 16–18.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Arnold, K., Bordoli, L., Kopp, J. and Schwede, T. (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, **22**, 195–201.
- Sadowski, M.I. and Jones, D.T. (2007) Benchmarking template selection and model quality assessment for high-resolution comparative modeling. *Proteins*, **69**, 476–485.
- Guex, N. and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.

35. van Gunsteren, W.F., Billeter, S.R., Eising, A., Hünenberger, P.H., Krüger, P., Mark, A.E., Scott, W.R.P. and Tironi, I.G. (1996) *Biomolecular Simulations: The GROMOS96 Manual and User Guide*. VdF Hochschulverlag ETHZ, Zürich.
36. Melo, F. and Feytmans, E. (1998) Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.*, **277**, 1141–1152.
37. Kopp, J. and Schwede, T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res.*, **32**, D230–D234.
38. Pieper, U., Eswar, N., Davis, F.P., Braberg, H., Madhusudhan, M.S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B.M., Eramian, D. *et al.* (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **34**, D291–D295.
39. Huang, H., Hu, Z.Z., Arighi, C.N. and Wu, C.H. (2007) Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. *Front Biosci.*, **12**, 5071–5088.
40. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A. (2007) UniProtKB/Swiss-Prot: The manually annotated section of the UniProt KnowledgeBase. *Methods Mol. Biol.*, **406**, 89–112.
41. Finn, R.D., Tate, J., Mistry, J., Coghill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
42. Mulder, N.J. and Apweiler, R. (2008) The InterPro database and tools for protein domain analysis. *Curr. Protoc. Bioinformatics*, **Chapter 2**, Unit 2.7.
43. Hartshorn, M.J. (2002) AstexViewer: a visualisation aid for structure-based drug design. *J. Comput. Aided Mol. Des.*, **16**, 871–881.
44. Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.*, **31**, 3381–3385.
45. Zhou, H. and Zhou, Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.
46. Wallner, B. and Elofsson, A. (2006) Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci.*, **15**, 900–913.
47. Benkert, P., Tosatto, S.C. and Schomburg, D. (2008) QMEAN: a comprehensive scoring function for model quality assessment. *Proteins*, **71**, 261–277.